A GEOMETRIC PERSPECTIVE OF STATISTICAL INFERENCE

Mitch Wilson
Brown Bag Presentation
GIDP in Applied Mathematics
University of Arizona

April 22, 2011

Layout

- Motivating Example: Austronesian Migrations
- Statistics Background
    - Relate to migration problem
- Introduction to Differential Geometry
    - Background
    - How it connects to the Statistics
- Current and Future Progress
- Conclusions

Austronesian Migration Patterns

- Indonesian island of New Guinea

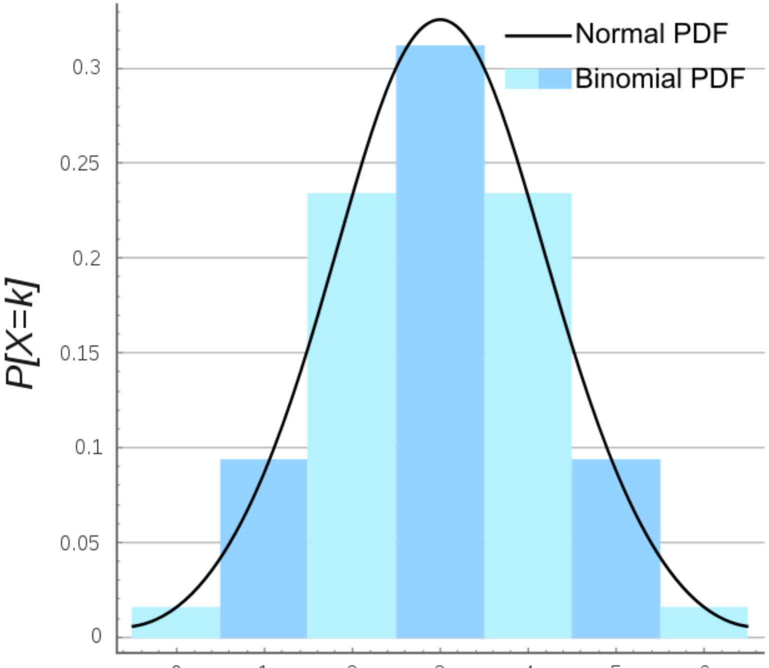
- Isolated for a very long time

Austronesian Migration Patterns

- Genetic samples from multiple islands
- Goal is to answer questions about mobility between islands, primarily:
  - Who traveled?
    - Western Indonesian, Oceania, Southeast Asia
  - When did these migrations occur?
    - Attribute to geographical or other factors?

Statistics Background

- Central Limit Theorem
- Score Functions and Maximum Likelihood Estimators (MLE’s)
- Fisher Information and the Cramer-Rao Bound



Central Limit Theorem

- Given a probability density function (pdf) $f(\mathbf{x}|\theta)$ with a given mean μ, the distribution of the value

$$\sqrt{n}(\bar{X} - \mu) \rightarrow W$$

  will have a Gaussian distribution with mean 0 and a covariance of ∑.
- This applies to all pdf's. with finite variance.

Central Limit Theorem

- Example: Exponential distribution $f(x|\lambda) = \lambda e^{-x\lambda}$
- Mean $1/\lambda$, variance $1/\lambda^2$
- Moment generating function (mgf):

$$M_x(t) = \left(1 - \frac{t}{\lambda}\right)^{-1}$$

$$E(e^{\lambda\sqrt{n}t(\bar{x}-\mu)}) = e^{-t\sqrt{n}} M_x\left(\frac{t}{\sqrt{n}}\right)^{n}$$

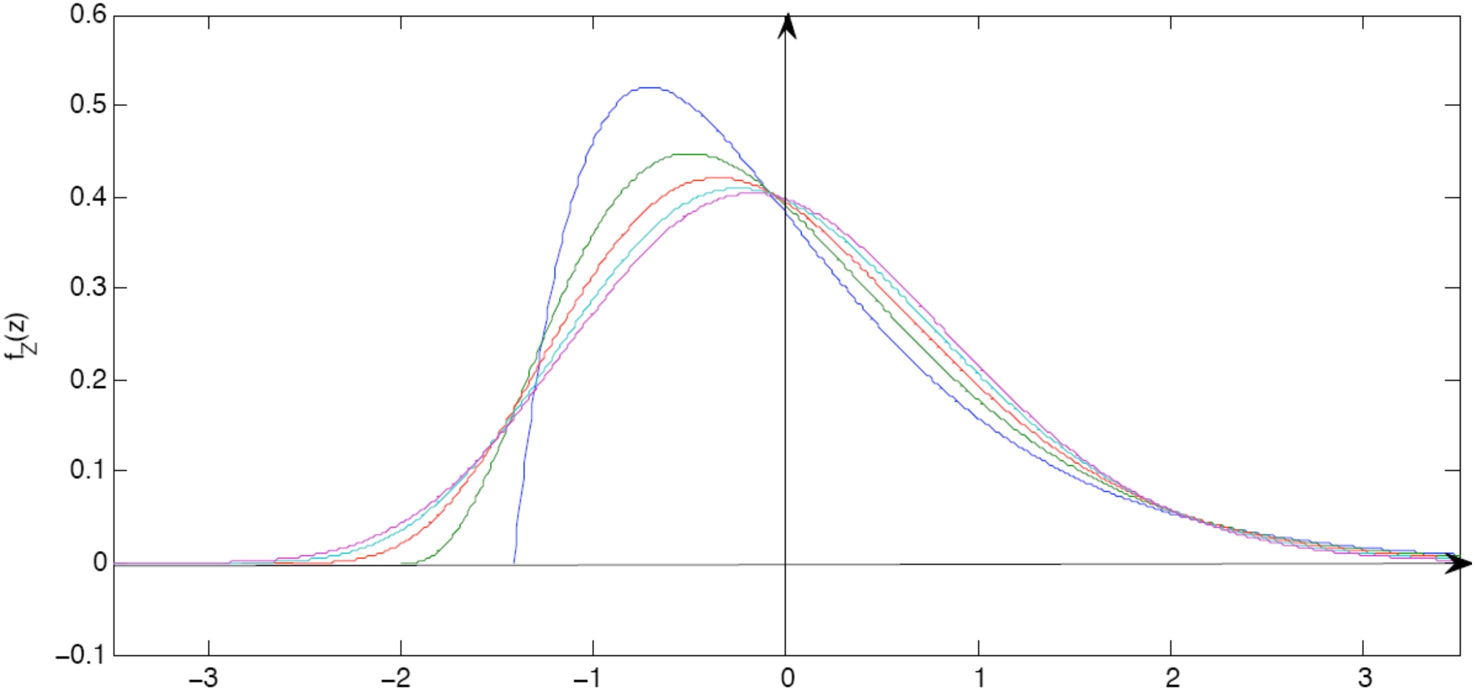
Central Limit Theorem

- Upon taking a logarithm, using a Taylor series, and retaking the exponential, we get

$$E(e^{\lambda\sqrt{n}t(\bar{x}-\mu)}) = e^{\frac{t^2}{2}+O(\frac{1}{\sqrt{n}})}$$

- As $n \to \infty$, this is the mgf of a normal distribution with mean 0 and variance 1.

Central Limit Theorem



Density of the standardized version of the sum of n independent exponential random variables for n = 2 (dark blue), 4 (green), 8 (red), 16 (light blue), and 32 (magenta).

The Search for an Estimator

- Family of distribution is assumed, but parameters θ are unknown
- Want an estimator, $\hat{\theta}$, with the following:
  - Low bias $E[\hat{\theta}] \approx \theta$
  - Low variance

Score Function

- ☐ Likelihood function

$$L(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- ☐ We define the score function as

$$\begin{aligned} \frac{d}{d\theta} \log f(x|\theta) &= \frac{d}{d\theta} \sum_{i=1}^{n} \log f(x_i|\theta) \\ &= \sum_{i=1}^{n} \frac{d}{d\theta} \log f(x_i|\theta) \end{aligned}$$

Maximum Likelihood Estimator (MLE)

- Most likely value of $L(\theta|x)$

$$\hat{\theta} = \max_{\theta} L(\theta|x)$$

- Can also be found by

$$\hat{\theta} = \{\theta : \frac{d}{d\theta} \log(L(\theta|x)) = 0\}$$

  - $\hat{\theta}$ is a critical point for the score function

MLE

- Looking at exponential distribution

$$\frac{d}{d\theta}\log L(\theta|x) = \frac{d}{d\theta}\log\prod_{i=1}^{n}\theta e^{-x_i\theta} = 0$$

$$\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \theta = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}} = \hat{\theta}$$

Fisher Information

- Variance of the score function

$$I(\theta) = E\left[\left(\frac{d}{d\theta}\log L(\theta|x)\right)^2 |\theta\right]$$

- An estimator $\hat{\theta}$ is unbiased if $E\left[\hat{\theta} - \theta\right] = 0$
- Using Cauchy-Schwartz, we obtain a useful inequality

Fisher Information

$$\frac{\partial}{\partial\theta}E[\hat{\theta}-\theta]=\int(\hat{\theta}-\theta)\frac{\partial f}{\partial\theta}dx-\int f\,dx=0$$

$$1=\int(\hat{\theta}-\theta)f\frac{\partial\log f}{\partial\theta}dx$$

$$=\int(\sqrt{f}(\hat{\theta}-\theta))(\sqrt{f}\frac{\partial\log f}{\partial\theta})dx$$

$$\leq\int f(\hat{\theta}-\theta)^2dx\int f\left(\frac{\partial\log f}{\partial\theta}\right)^2dx$$

Cramer-Rao Bound

$$\sigma_{\hat{\theta}}^2 \geq \frac{1}{I(\theta)}$$

- The more information we have about the distribution, the narrower the variance, and vice-versa.
- If the pdf belongs to an exponential family, then the lower bound becomes an inequality.

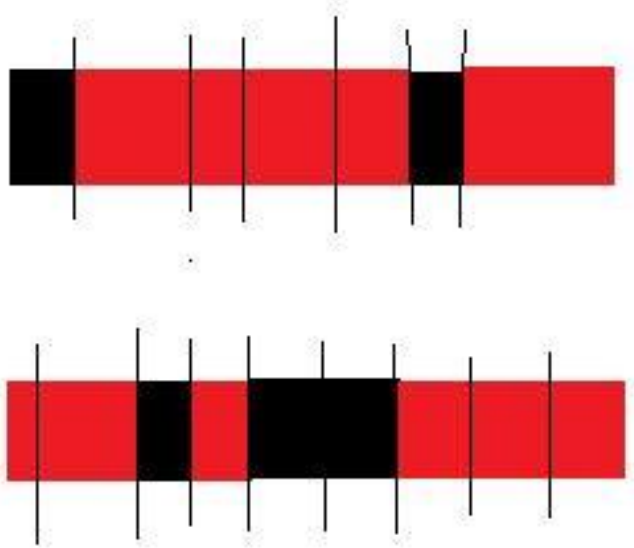
$$f(x \mid \theta) = h(x)\exp(g(\theta)T(x) - A(\theta))$$

Asymptotics on MLE's

- Why are MLE's good to use?
- As $n \to \infty$, $(\hat{\theta} - \theta) \to 0$, so the MLE will converge to the true mean
- The value $\sqrt{n}(\hat{\theta} - \theta)$ will have a mean of 0 and a variance of $\frac{1}{I(\theta)}$ as $n \to \infty$, so the Cramer-Rao bound will approach equality asymptotically.

Integration with Migration Model

- Multiple segments of native and foreigner DNA
  - Hard to distinguish particular lengths
- Can be viewed as geometric distribution with parameter π, the probability of the next strand containing native DNA
- Approximate as exponential distribution on a larger scale with parameter τ, the time of contact in generations.



Estimating DNA Length

- Given n strands of DNA, each with length l_i and total length $$L = \sum_{i=1}^{n} l_i,$$ we can use the Laplace transform $E[e^{sL}]$ to estimate the parameter λ_n representing the total length of the native strands

Estimating DNA length

- By approximating our geometric distribution as an exponential one with parameter τ, for a single element we find

$$E[e^{sl}] = \int_0^\infty e^{sl} \tau e^{-\tau l} dl = \frac{\tau}{\tau - s} \text{ if } s < \tau.$$

- Thus, for our whole sequence

$$E[e^{sL}] = \sum_{n=1}^{\infty} E[E[e^{sL}]|N = n]P(N = n)$$

Estimating DNA length

$$E[e^{sL}] = \sum_{n=1}^{\infty} \left(\frac{\tau}{\tau - s} \right)^n \pi^{n-1}(1 - \pi)$$

$$\lambda_n = \frac{\dfrac{\tau}{\tau - s}(1-\pi)}{1-\dfrac{\tau\pi}{\tau - s}} = \frac{\tau(1-\pi)}{\tau(1-\pi)-s}$$

$$\lambda_n = \tau\pi \text{ and } \lambda_f = \tau(1 - \pi)$$

MLE and Fisher Information

- Using the MLE, we obtain

$$\hat{\tau} = \lambda_f + \lambda_n \text{ and } \hat{\pi} = \frac{\lambda_n}{\lambda_n + \lambda_f}$$

  - These agree with our estimates from earlier.

- In multiple dimensions, the information has a covariant matrix with indices

$$I(\theta)_{i,j} = E\left[\left(\frac{\partial}{\partial \theta_i} \log f(\mathbf{X}|\theta)\right)\left(\frac{\partial}{\partial \theta_j} \log f(\mathbf{X}|\theta)\right)|\theta\right]$$

Migration Fisher Information

$$\log(L(l|\lambda)) = \log(\prod_{i=1}^{n} \lambda_n e^{-l\lambda_n} \prod_{i=1}^{n} \lambda_f e^{-l\lambda_f})$$

- Off-diagonal entries are zero
- The diagonal entries are that of the variance

$$\tilde{I}(\lambda_n, \lambda_f) = LK \begin{pmatrix} \frac{1}{\lambda_n^2} & 0 \\ 0 & \frac{1}{\lambda_f^2} \end{pmatrix}$$

Migration Fisher Information

- Let $g : (\lambda_n, \lambda_f) \rightarrow (\pi, \tau)$
- Define the Jacobian on g:

$$J_g(\tau, \pi) = \left(\begin{array}{cc} \frac{\partial \lambda_n}{\partial \tau} & \frac{\partial \lambda_n}{\partial \pi} \\ \frac{\partial \lambda_f}{\partial \tau} & \frac{\partial \lambda_f}{\partial \pi} \end{array} \right) = \left(\begin{array}{cc} \pi & \tau \\ 1 - \pi & -\tau \end{array} \right)$$

- The Fisher Information on this new space is

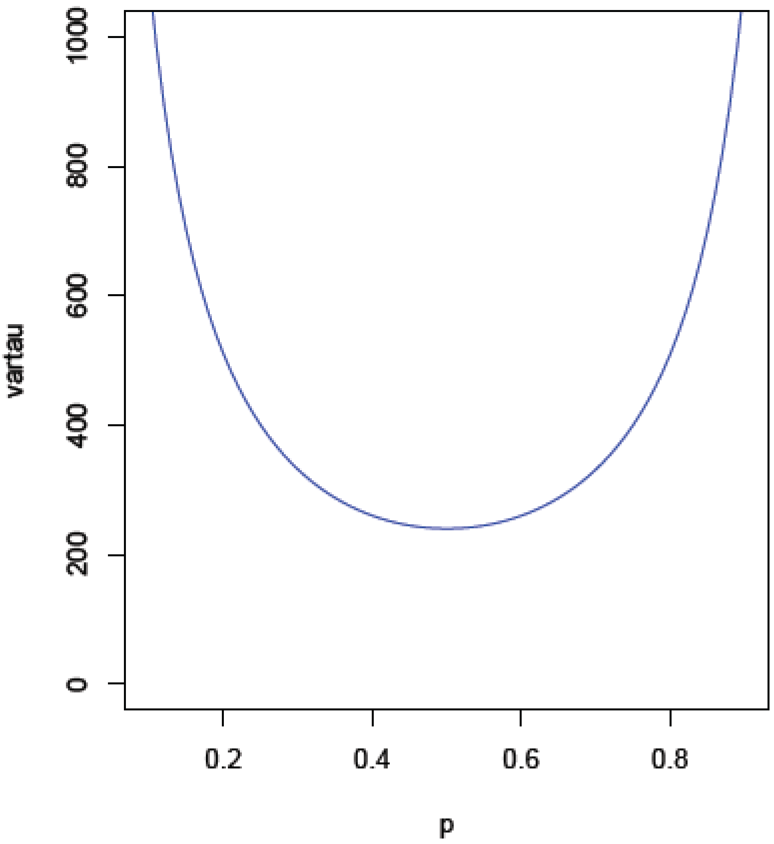
$$I(\tau, \pi) = J_g(\tau, \pi)^T \tilde{I}(g^{-1}(\tau, \pi)) J_g(\tau, \pi)$$

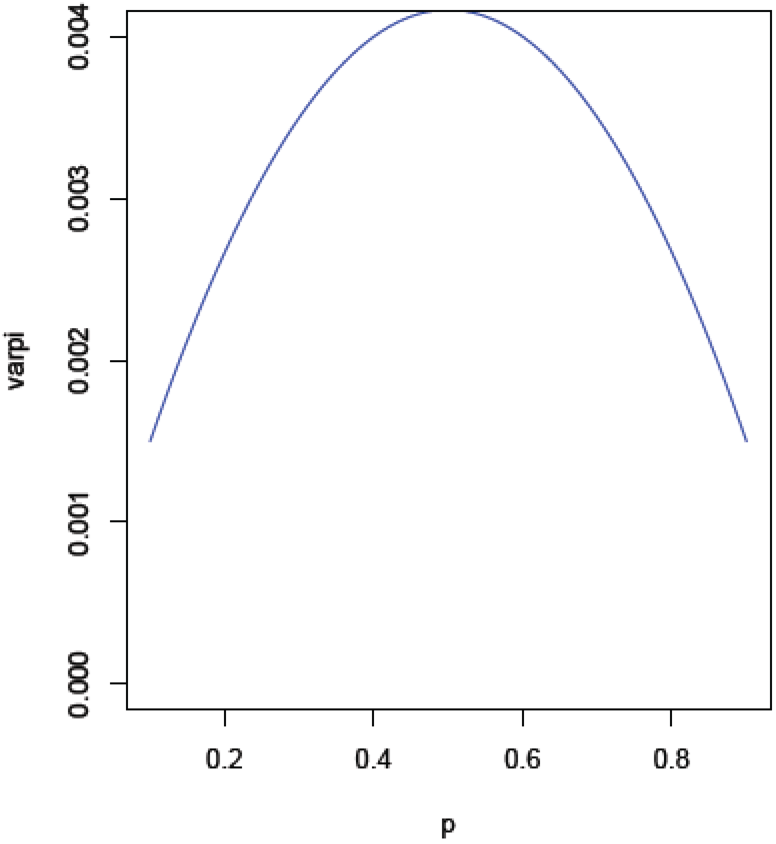
Migration Fisher Information and Covariance

$$I(\tau,\pi) \approx \frac{L\pi(1-\pi)}{\tau} \begin{pmatrix} 2 & \frac{\tau(1-2\pi)}{\pi(1-\pi)} \\ \frac{\tau(1-2\pi)}{\pi(1-\pi)} & \frac{\tau^2((1-\pi)^2+\pi^2)}{\pi^2(1-\pi)^2} \end{pmatrix}$$

$$\mathrm{Cov}(\hat{\tau},\hat{\pi}) = \frac{1}{L} \begin{pmatrix} \frac{\tau((1-\pi)^2+\pi^2)}{\pi(1-\pi)} & -(1-2\pi) \\ -(1-2\pi) & \frac{2\pi(1-\pi)}{\tau} \end{pmatrix}$$

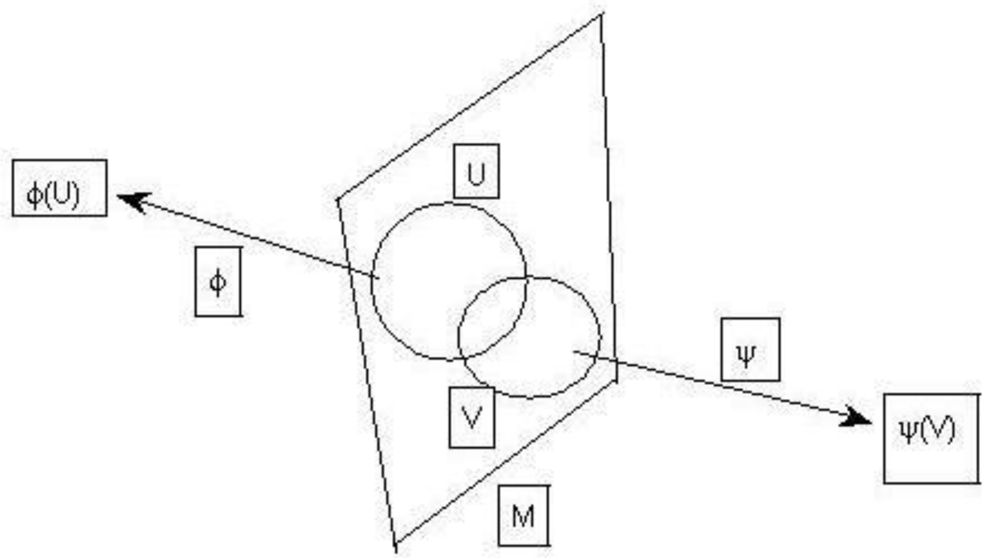
Variance of Estimators





Geometry Background

- Manifold
- Tangent Spaces and Submanifolds



Riemannian Metric

- (Semi-positive definite) inner product on tangent spaces $\left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots\right)$
  - can also be viewed as a tensor.

- Its indices are given by, for a point p,

$$g_{i,j}(p) = g_p\left(\left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p\right)$$

Riemannian Manifold

- If D and D’ are tangent vectors, and the following properties hold:
  - Linearity
  - Symmetry
  - Positive Definiteness
- then we have a Riemannian metric
- A Manifold with a Riemannian Metric is (naturally) called the Riemannian Manifold

The Covariant Derivative

- The change of a tangent space in one direction, with respect to another direction, can be given by

$$\nabla_{\partial i}\partial j = \Gamma_{ij}^{k}\partial_{k}$$

where Γ_{ij}^{k} is the connection coefficient between neighboring tangent spaces.

Relating Geometry to Statistics

- Work of Amari, others
- A manifold in $\Re^d$ with $g(\theta_i, \theta_j) = I(\theta)_{i,j}$ relates the Fisher information to the Riemannian metric
- Relate information in different spaces with Jacobian

Current Work

- The two-migration model
- Two population parameters π_1, π_2
- Two time parameters τ_1, τ_2
- Hypothesis testing becomes singular as parameters reach extreme cases: $\pi_1 \text{ or } \pi_2 = 0$

$$\tau_1 = \tau_2$$

- Look at it from Geometry perspective

Current Work

- Find g: $\Re^4$ to $\Re^6$

$$\lambda_1, \lambda_2, \lambda_3, p_{12}, p_{13}, p_{23}$$

- Find Fisher Information on this less complex submanifold.
- Revert data to original submanifold, make inferences based on these results.

Expectation-Maximization

- Mixtures of distributions are subject to singularities
  - Calculating things like the MLE may not be possible
- Marriot, et al. (2009) propose introducing a penalty function to discourage certain parameters from driving towards more troubled regions.
- The idea: use an iterative process that will converge to the optimal set of parameters by choosing some values ad hoc at the beginning.

The E-M Algorithm

- Suppose weights α_1, α_2, etc. are unknown and other parameters β_1, β_2, etc. are known
- Step 1: Make an intuitive guess for the values of α
- Step 2: Compute the initial values of β which will maximize the likelihood function (with perhaps a penalty added)
- Step 3: (E-step) Determine intermediate weights based on the values of α and the distributions. These are the conditional expectations.

The E-M Algorithm

- Step 3 (cont’d): (M-step) Update the values of α and β by maximizing the weighted likelihood function.
- Step 4: Calculate the likelihood after every iteration. Repeat the process until the parameters no longer change.

Future Work

- Continue work on two-migration model
  - Use E-M Test to minimize the dangers of going to extreme cases
  - Expand to regions where the DNA differences are not as well isolated
- Can apply to other phenomena
  - Queuing processes
  - Neural Networks
  - Finite state Markov processes

Conclusions

- By looking at statistical problems from a Geometry point of view, we can gain new insight and other ways to evaluate the data that we have.

- I would like to give a big thank you to Dr. Watkins for introducing me to this interesting topic. I would also like to thank the University of Arizona GIDP in Applied Mathematics for making this research possible.

References

- Amari, et al; "Differential Geometry in Statistical Inference", Institute of Mathematical Sciences, 1987
- Amari; Nagaoka; "Methods of Information Geometry", AMS, 1993 (translated 2000)
- Cassela; Berger; "Statistical Inference," Duxbury, 2002
- Hicks, Noel; "Notes on Differential Geometry." Redwood Burn Limited, 1971
- Karafet, et al; "Major East-West Underlies Y Chromosome Stratification across Indonesia," Oxford Journals, Mol. Biol. Evol. 27(8), 2010
- Li; Chen; Marriot; "Non-Finite Fisher Information and Homogeneity: an E-M approach", Biometrika (2009), 96 2 p. 411-426
- Watkins, Joseph; Math 363/466/564/566 notes

Questions?

- Contact information:

Mitch Wilson

mjw@math.arizona.edu

math.arizona.edu/~mjw

Dr. Joseph Watkins

jwatkins@math.arizona.edu

Math.arizona.edu/~jwatkins